# DEMOCRATIC GOVERNANCE OF FRONTIER AI

## A close reading of OpenAI's June 2026 federal framework, provision by provision, for the question Congress actually has to answer

MARKUP | June 18, 2026

By Fatima Faisal Khan, Institute for Security + Technology

---

On June 2, 2026, OpenAI published "Democratic Governance of Frontier AI: A blueprint for a federal framework," a proposal for how the United States should govern its most capable AI systems. The blueprint arrived the day after the White House's executive order on Promoting Advanced Artificial Intelligence Innovation and Security, in the same week that Sam Altman met congressional leaders on the Hill, and alongside the bipartisan Great American AI Act draft from Representatives Obernolte and Trahan. In a moment when Washington was searching for a federal approach, a frontier developer offered a full framework. It proposes three things:

1. A "reverse federalism" that codifies recent and pending state laws, including California's SB 53, New York's RAISE Act, and Illinois's SB 315, and then preempts the states.
2. A strengthened Center for AI Standards and Innovation (CAISI) as the government's premier institution for evaluating frontier models and measuring recursive self-improvement
3. A whole-of-government resilience strategy across compute, biodefense, and cybersecurity

In much of the commentary featured in Politico, The Hill, Mashable, and others immediately following the release of the framework, most asked why OpenAI wrote this. Rather than retread that well-covered ground, I focused on the question: what would be required to make the framework work? There are three core challenges that policymakers would face were they try to implement the framework as it stands.

First, the blueprint reads as a list of obligations but is really a sequence with unstated prerequisites. Its most important requirements—the severe-risk evaluations, comparable transparency reports, third-party audits, and the procurement gate—rely on

methodologies and deliverables that do not yet exist. These include the evaluation methodologies, the metrics for recursive self-improvement and loss of control, and a certified assessor ecosystem. It compresses what will be a multi-year build into single sentences, which makes the real hazard a sequencing one. Enacting the fast structural pieces, particularly the federal preemption that displaces state AI laws, before the slow substantive ones exist risks a gap in which operative state law is gone and the federal regime cannot yet function.

Second, the framework relies on measurement and definitions that do not exist. Recursive self-improvement, loss of control, and even cyber, chemical, biological, radiological, and nuclear (CBRN) capability lack agreed evaluation methods, and the document's own fix is to have CAISI develop them later, in conjunction with the firms being measured. As a result, key terms remain unclear. "Severe risk" carries the evaluation, whistleblower, and liability provisions without a definition. The phrase "critical safety incident" brings to mind issues with implementing the Cyber Incident Reporting for Critical Infrastructure Act of 2022: four years after passage, we have yet to reach a settled definition of a reportable "critical" incident. Pinning these terms down is what determines whether anything is enforceable.

Third, the framework's binding force depends on choices it leaves to implementation. The centerpiece is a mandatory pre-release CAISI evaluation, but the evaluator cannot block deployment, and a deadline lets developers ship if CAISI does not finish in time. The gate therefore binds in proportion to CAISI's capacity, so the agency would need the staffing and tooling to keep pace with every frontier release before the requirement carries real force.. A separate design choice is the question of who runs the assessments. The framework assumes a private assessor ecosystem stood up from scratch, though at the current scale of roughly five frontier labs a government-run assessment could be faster and more independent. Policymakers need to decide which route to build and fund. The hardest piece to pin down is the backstop. Liability for severe harms with no blanket safe harbor is what would give the other obligations consequences, yet the blueprint states it as a single principle, leaving the work of turning it into a workable standard as the thing that determines whether the rest of the regime has force.

This is among the first comprehensive frameworks any frontier developer has put before policymakers, and its main contribution may be to show the scale of the task it leaves to the government. The framework presents a set of obligations, but reads more like a sequence in which the faster structural elements, preemption above all, are specified far more fully than the measurement methods, threshold definitions, and assessor

institutions on which everything else depends. Most of what would make the regime function has yet to be built, and much of that building falls to policymakers rather than to the company making the proposal.

This markup is the first in a series examining how frontier developers propose to govern themselves. Eight days after the OpenAI document, Anthropic published its Advanced AI Framework alongside Dario Amodei's essay Policy on the AI Exponential, proposing an Federal Aviation Administration (FAA)-style testing regime in which the government can block or reverse deployments that fail. That is a sharply different answer to the question this markup turns on, since OpenAI's evaluator cannot block deployment, and it is the framework I will take up next.

What follows below is my provision-by-provision markup, from the preamble through the conclusion of OpenAI's federal framework, testing each recommendation against what it would take to build and whether it would hold.

---

Key Analysis sections appear in teal boxes

*The countries that successfully harness artificial intelligence will shape the scientific, economic, and geopolitical trajectory of the 21st century. AI can accelerate scientific discovery, expand economic opportunity, strengthen national security, and help solve problems that once seemed unsolvable. At the same time, increasingly capable AI systems are beginning to demonstrate abilities that raise concerns about cyber offense, biological misuse, autonomy, alignment, and other threats to national security. We also see early signs of recursive self-improvement in today's systems: where AI development is itself accelerated by AI. We expect this to increase competitive pressures among developers and nations, and create governance challenges that existing institutions are not equipped to address. As RSI emerges, societies will need ways to shape the trajectory of AI development and ensure that it serves human interests.*

**Analysis**

The framework introduces RSI in its opening paragraph. RSI acts as the document's "spine" and frames the claims that follow. It shifts AI policy from the register of consumer protection and competition into the register of national security—one in which the state is expected to do the heavy lifting, while incumbents remain indispensable partners rather than regulated parties. The document wants institutions built to track RSI, but there is no agreed way to measure RSI today, a gap the document itself acknowledges later when it tasks CAISI with developing the methodology. An obligation to monitor something the field cannot yet define cannot bind anyone until the measurement exists, so the premise is being asked to carry institutional weight that the underlying science does not yet support.

*Democracies are uniquely positioned to ensure that powerful AI is developed responsibly by coupling innovation with public accountability, transparency, independent oversight, and the ability to course-correct through representative government. But effective governance requires visibility into how frontier capabilities are evolving, how AI is impacting national security, and whether changing risk profiles warrant additional domestic safeguards, international coordination, or other precautionary measures. Building that understanding necessitates creating institutions capable of evaluating frontier AI, monitoring how capabilities evolve, and providing policymakers with reliable information.*

*As AI becomes increasingly important, democratic governments—not private companies acting alone—must ultimately determine the rules, safeguards, and accountability mechanisms. Our view is that decisions about the pace of AI innovation should not be left to any one lab, company, or special interest group. Instead, these choices should be made through democratic processes and informed by a robust understanding of frontier capabilities, risk mitigation measures, societal resilience, and geopolitical considerations.*

**Analysis**

Policymakers will need to develop their own endogenous policy development capacity to fulfill this principle. Merely performing a ministerial function of ratifying private companies' proposals would not meet the spirit of a democratic process in which society charts its own course.

*In this context, we believe that the United States is particularly well positioned to help shape global governance. The US federal government possesses unique capabilities that no private company can replicate, including access to classified intelligence, expertise in cybersecurity and chemical, biological, radiological, and nuclear CBRN defense, secure computing environments, and the ability to coordinate with international partners. And many building blocks of a frontier safety framework already exist in the US today. Frontier developers have already adopted White House voluntary commitments and partnered with the Center for AI Standards and Innovation (CAISI) for pre-deployment evaluations. US companies have also coordinated internationally, signing onto the European Union's AI Act Code of Practice and partnering with the United Kingdom's AI Security Institute. States have started developing harmonized approaches to frontier AI governance that include California's SB 53, New York's RAISE Act, and Illinois's SB 315, and the White House's recent executive order on Promoting Advanced Artificial Intelligence Innovation and Security is an important step forward. The US federal government must now build on that foundation and create a durable federal framework capable of evolving alongside the technology itself.*

**Analysis**

The unique capabilities claim calls for the government to carry out the evaluation, not for industry to play a role in co-developing the evaluator. Co-development only becomes the answer once you add what the clause leaves out: a frontier evaluation also needs model access, frontier methods, and the developers' own talent, none of which the government currently holds. The government could acquire them and evaluate models on its own. The co-development of CAISI's methodologies that the framework proposes later is therefore a design choice for closing that gap quickly, not a consequence of the capabilities named here. The capability claim does a second kind of feasibility work too. Classified intelligence, CBRN expertise, and secure compute are real government assets, but bringing them to bear on model evaluation requires interagency mechanisms, clearances, and data-sharing arrangements that do not currently exist in usable form. These dissemination and classification

challenges are ones policymakers will need to address, and they may well dwarf the challenge of gathering the intelligence in the first place.

**Analysis**

While the June 2 White House Executive Order focuses on voluntary preclearance, this blueprint calls for a mandatory pre-release evaluation. The building blocks listed (voluntary commitments, the EU Code, several state laws, CAISI) do not share definitions, thresholds, or reporting formats, so harmonizing them into one coherent federal framework is itself a substantial drafting and standard-setting project, rather than a starting point that can simply be assumed as a solid foundation.

*If artificial general intelligence is going to benefit all of humanity, the world needs more than voluntary commitments, individual company policies, and isolated regulatory interventions. It needs harmonized legal frameworks and durable institutions capable of adapting as technology advances. That framework should:*

*Address frontier risks to national security and public safety. The primary goal of any frontier safety framework should be to mitigate the most severe risks posed by advanced general-purpose AI systems. These include risks related to cyber and CBRN threats, RSI progress, and loss-of-control scenarios that could result in catastrophic outcomes.*

**Analysis**

This sentence defines the scope of "safety" for the whole document, and the definition is narrow. It is the catastrophic tail, which includes cyber, CBRN, RSI, and loss-of-control. Privacy, discrimination, fraud, labor displacement, consumer harms (i.e., the harms that drive most AI regulation and that impose ongoing compliance costs on a product business) are absent.

Narrowing "safety" to the tail is the single most consequential framing choice in the document. It is defensible (in that this is where catastrophe lives), and it is also convenient (it routes the burden to government and away from frontier model producers).

A narrow, tail-focused mandate is, in principle, the more tractable kind to administer, since there are fewer high-stakes risk categories to evaluate. There are two catches. Two of the four named domains, loss of control and RSI, are precisely the ones with no accepted measurement, so the framework's hardest operational problems sit at its very core rather than at its edges. The same vagueness recurs in the term "severe risk" itself, which the document never defines and which several later provisions (evaluations, whistleblower protection, liability) all lean on. Moreover, under this framework, privacy, discrimination, fraud, and the other near-term harms are left to whatever regime falls outside the scope of preemption, and the document does not say what that regime is.

- *Advance democratic governance. Decisions about how society manages frontier AI risks should be made through representative government, not by private companies acting alone. Frontier safety governance should reflect the strengths of free societies: transparency, public accountability, independent oversight, and the rule of law.*
- *Promote transparency. Governments, researchers, businesses, and the public need reliable information about how frontier AI is developed, evaluated, and deployed. Transparency creates*

*accountability, supports independent scrutiny, and helps ensure that policy decisions are informed by evidence.*

- *Protect innovation. A frontier safety framework should focus on the highest-consequence risks without creating unnecessary barriers for startups, researchers, and developers building on top of frontier capabilities. It should reduce risk without locking today's industry structure into law.*

**Analysis**

This is the explicit rebuttal to the barrier-to-entry critique, as narrow targeting is less burdensome to small firms than a flat regime. However, the protection is asserted rather than engineered. The audit and pre-deployment evaluation requirements still function as a tollbooth at the moment a startup tries to cross into the frontier tier. Whether this principle holds depends entirely on where the (undefined) threshold sits and who pays for compliance.

Until the threshold and the cost incidence are written down, it remains unclear whether this principle protects innovation or entrenches incumbents, a dichotomy that the framework leaves for future action.

- *Build adaptive institutions. AI is advancing rapidly, and frontier governance must be capable of evolving alongside the technology. Policymakers should create institutions that can learn, experiment, incorporate new evidence, and update standards over time.*

**Analysis**

The corollary for “adaptive” institutions is “discretionary,” and a body that updates its own standards is continuously open to influence by the most technically capable party before it.

Building adaptive institutions is the right goal, but the framework’s core asks—the evaluation, audit, and reporting obligations on companies—are binding requirements rather than voluntary standards, and adaptivity works differently for the two. A body that revises binding requirements cannot rewrite them at will. Binding rules come with the process the Administrative Procedure Act exists to impose: notice and comment, a reasoned basis for each change, and some insulation from the regulated parties. The blueprint leaves that process out and the omission cuts both ways, since open public consultation is what keeps revision from sliding into capture, and predictability is what lets industry plan against the rules rather than only react to them.

*Implementing a frontier safety framework will require action at multiple levels of government, as well as international cooperation. States can continue serving as laboratories of democracy by developing harmonized frontier safety laws. The US federal government should build on that foundation by creating institutions capable of evaluating frontier systems, identifying emerging risks, and informing decisions about how AI should be governed. And as AI becomes more powerful, policymakers will need a whole-of-government plan to build broader societal resilience. This blueprint outlines a three-part strategy for achieving those goals: 1) building a national framework that leverages the emerging consensus reflected in state frontier safety laws; 2) strengthening CAISI as the US federal government's primary institution for frontier AI safety; and, 3) mobilizing a broader resilience plan across government to address the national security and public safety challenges posed by frontier AI.*

## 1. Building a national framework through reverse federalism

*States have been valuable laboratories for AI policy, helping to develop and test many of the ideas that now form an emerging consensus on frontier safety. As AI becomes increasingly capable, however, the most important frontier safety challenges will be national—and often international—in scope. Policymakers should build on that foundation and turn today's emerging consensus into a comprehensive federal framework. This approach, which we call reverse federalism, allowed states to develop and refine common legal frameworks first, creating models that Congress should now adopt at the national level. California's SB 53, New York's RAISE Act, and Illinois's SB 315 demonstrate that a meaningful consensus has emerged around the core elements of frontier AI governance. These frameworks share common requirements, align with international approaches, and provide a practical starting point for federal legislation. Policymakers should build on this consensus by establishing a national frontier safety framework that provides both robust safeguards and regulatory certainty. At a minimum, that framework should include:*

**Analysis**

Sequencing is the acute risk in evaluating whether this works as designed. Preemption only delivers the "regulatory certainty" the document promises if the federal framework is actually complete and operational at the moment state law is switched off. If Congress preempts before CAISI's standards and evaluation methods exist, which according to this document's logic is still years from now, the result is a gap: state law is gone and the federal replacement is not yet functional. The gap matters because the state laws are operative, not placeholders: SB 53 is already in force with its own definitions and penalties, which Congress could lift entirely for overlapping requirements. What it cannot lift is the RSI and loss-of-control measurement the framework adds on top, as that piece does not yet exist.

The phrase "the same frontier safety risks," which scopes the preemption, has to be defined with real precision in statute, or it will be defined for Congress by litigation. It is the kind of term that invites years of boundary disputes over what constitutes frontier safety versus ordinary consumer protection.

- *Severe risk evaluations and mitigations. Companies should evaluate frontier capabilities for risks related to cyber, CBRN, loss of control, misalignment, and progress towards RSI; implement appropriate safeguards; and explain why any residual risks are appropriately managed. Risk assessments and mitigations should be tailored to the model's deployment context.*
- *Transparency requirements. Companies should publish public frontier safety frameworks and transparency reports describing how they evaluate severe risks, implement safeguards, make deployment decisions, and responsibly track progress towards RSI, with appropriate redactions to protect security, trade secrets, and proprietary information.*

**Analysis**

Comparable, published disclosure is what lets researchers, government, and competitors scrutinize how a model was evaluated and what was decided, which is the factual basis any later oversight depends on. It is also the rare provision that can be stood up quickly, because templates already exist.

Transparency requirements are the most build-ready item in the document for that reason, and the practical move is to point policymakers at the artifacts that already exist, such as the published frontier safety frameworks and transparency reports from the major labs, plus the disclosure schemas in SB 53 and the EU Code of Practice. These give Congress a concrete starting schema to adopt rather than invent.

As written, a frontier lab would author the framework, write the report, and decide what to withhold under "appropriate redactions to protect security, trade secrets, and proprietary information," categories broad enough to absorb almost anything. The annual independent audit does not address this, since it tests compliance with the lab's own framework, rather than the report's completeness or the redactions. To operate, the provision needs a defined content schema (what fields, what granularity), a cadence (per model, annual, or on material change), a filing venue, and an adjudicator for contested redactions, with a confidential unredacted version filed with a regulator so completeness can be checked. The framework names the deliverable and leaves all of that to be specified.

- *Independent assessment and auditing. Large frontier developers should annually retain an independent third party to audit compliance with frontier safety requirements, including implementation of the developer's frontier AI framework, internal controls, and governance structures. These audits should be underpinned by a set of common standards that allow for interoperable audits across jurisdictions.*

**Analysis**

The standard auditor-independence problem applies, since the auditor is hired and paid by the audited. To address this, standard fixes, such as rotation and independent selection, would be needed.

The deeper question is who the auditors even are. The provision assumes a pool of third parties qualified to audit frontier AI, and that pool does not meaningfully exist today. As a result, the framework elsewhere has to ask CAISI to certify assessors and help build the assessment ecosystem. The audit requirement quietly depends on a certification and ecosystem deliverable that has not yet been built.

This lack of a certification and ecosystem deliverable raises a real design alternative the document skips. Third-party assessor regimes (like the FedRAMP 3PAO model) are useful when the volume of assessments is too high for the government to handle directly. With only a handful of frontier labs, the volume is low, so it is fair to ask why this defaults to a third-party ecosystem that must be created from scratch rather than to second-party, government-conducted assessment, which could be faster to stand up and easier to keep independent. At this scale, "build a certified private ecosystem" is a choice, not a necessity, and Congress should weigh the government assessor option on its merits.

"Interoperable standards across jurisdictions" is double-edged: it reduces duplicate burden, and it also makes it easier for industry to shape a single set of standards, increasing the risk of regulatory capture. It also presupposes common audit standards that do not yet exist, another dependency on work not yet done.

- *Critical safety incident detection and reporting. Companies should report critical safety incidents involving deployed models, including incidents related to risks covered by the frontier safety framework, dangerous model behavior, or unauthorized access to sensitive model weights.*

**Analysis**

The Obernolte-Trahan discussion draft of the Great American AI Act also includes incident reporting, making this the area of broadest cross-document agreement.

The hard part is defining the trigger, and there is direct precedent for how hard that is. CIRCIA was enacted in 2022 and, four years on, still lacks a settled definition of a reportable cyber incident. In the AI case, the concepts of "critical safety incident" and "dangerous model behavior" could provoke the same multi-year definitional fight, and that definition, not the reporting duty itself, decides whether the regime captures real events or collapses into noise or nothing. The provision needs a defined incident threshold, a reporting clock, a named recipient, and a confidentiality regime to function.

Note that the framework presumes an observable, controlled deployment and does not map onto open-weight release. Once weights are out, the developer cannot observe incidents, so the regime structurally exempts open models, which is a coverage hole worth naming explicitly.

- *Model weight security requirements. Companies should implement cybersecurity and insider-threat protections to secure unreleased model weights.*
- *Whistleblower protections. Employees should be protected from retaliation when reporting credible concerns about severe risks, safety failures, critical safety incidents, or violations of law to company leadership, regulators, or other appropriate authorities.*

**Analysis**

This proposal is one of the easiest to legislate because whistleblower statutes have well-worn templates such as Sarbanes-Oxley, Dodd-Frank, and the AI-specific whistleblower protections California already enacted in SB 53. The operational choices are scope (which disclosures and to whom: leadership, a regulator, or the public); the standard of protection (anti-retaliation with burden-shifting); and the enforcement venue, all of which have models elsewhere in statute that Congress can draw from for AI-specific applications.

Whistleblower protections outlined in this framework are prompted by reports of "severe risks, safety failures, critical safety incidents, or violations of law" and "severe risk." However, none of those categories are defined with the precision a retaliation case would require. A whistleblower regime is only as enforceable as the conduct it protects reporting on, so the undefined threshold weakens this provision the same way it weakens the evaluation and liability provisions that rely on the same definitions and precision.

- *Meaningful accountability mechanisms. Companies should face enforceable consequences for failing to comply with transparency, reporting, and safety obligations.*

**Analysis**

This framework hinges accountability on “severe harms,” whereas other liability regimes typically hinge on conduct, a standard of care such as negligence or recklessness, rather than merely on the magnitude of the harm. Pegging liability to harm severity leaves significant harms without recourse and imports the same narrow scope choice that I wrote about in discussion of Principle 1. It is worth asking directly how this compares to existing liability regimes.

“No blanket safe harbors” is therefore a stance rather than a standard. The distinction worth drawing is that auditing and disclosure are the framework’s ‘detection layer,’ the parts that surface whether a developer met its obligations. Meanwhile, this provision is the ‘consequence layer,’ the part that decides what follows if a developer did not meet its obligations. Detection without a defined consequence does not bind, so this least-developed provision is the one that gives the rest of the provisions their teeth. To operate, it has to specify the standard of care, who has standing, and whether complying with the framework is a safe harbor or merely evidence of due care.

*Liability frameworks should preserve accountability for severe harms and should not provide blanket safe harbors from responsibility. The requirements reflected in SB 53, RAISE, and SB 315 should serve as the foundation for federal frontier safety legislation—not its endpoint. Policymakers should build beyond them by establishing a formal role for CAISI, strengthening federal evaluation and assessment capabilities, and advancing the broader resilience measures described below. With this comprehensive federal framework in place, policymakers should also preempt state laws that seek to regulate the same frontier safety risks, creating a single national framework that combines strong safeguards with regulatory certainty. States should continue serving as laboratories of democracy in areas beyond frontier safety, including youth protection, electricity and environmental policy, and AI education and literacy.*

## 1. Strengthening safety through strong institutions

*As AI becomes increasingly capable, the US will need a trusted institution responsible for evaluating frontier AI, monitoring emerging risks, and providing policymakers with independent technical advice. RSI exacerbates the fundamental governance question of whether humans can retain the ability to understand, guide, and shape the trajectory of advanced AI: making it potentially the most consequential frontier safety issue of the coming decade. Yet policymakers currently have limited visibility into RSI progress, whether safeguards are keeping pace, or what indicators should inform future policy decisions. An institution like CAISI can help close that gap. Policymakers should strengthen CAISI and build it into the world's premier institution for frontier AI evaluation, standards development, independent assessment certification, and coordination across government and with international partners. Its mission should be to provide the information and analysis that policymakers need to make informed decisions about national security, international coordination, the adequacy of safeguards, and the pace of AI development. Particular priority should be given to understanding progress toward RSI, developing reliable measurements of that progress, and ensuring that governments have the information needed to respond. As capabilities advance, CAISI should remain flexible and adaptable, with the ability to take on new responsibilities as emerging risks and governance needs become clearer.*

**Analysis**

The complication is that the CAISI’s evaluation methods are meant to be developed together with industry. Some industry input is unavoidable, because the relevant technical expertise sits largely inside the labs. The question is not whether to include it, but how to structure its entry: published methodologies, independent validation, and final authorship that stays with the public body rather than

with the firms being measured. Without that structure, an authoritative institution running on industry-set standards lends the credibility of government judgment to what is effectively industry preference, which is a worse position than a more modest body that does not carry that imprimatur.

“Independent assessment certification”

Certification implies a third-party model, in which CAISI accredits private assessors rather than assessing models itself. That model, the FedRAMP 3PAO approach, exists to handle assessment volume too large for the government to manage directly. As mentioned previously, at the current scale of roughly five frontier developers, that rationale is thin. However, independent does not have to mean third-party; a second-party model in which government conducts the assessments is also independent of the developer, and at this scale it may be faster to stand up and harder to capture, since it avoids the conflict of an assessor paid by the firm it assesses and keeps the work close to the classified data the document says only government holds. Volume could grow as model versions and internal deployments multiply, so a hybrid is reasonable, but the blueprint should justify the third-party default rather than treat it as a given.

What decides the outcome here is the design of the institution, not whether it exists. These design details are not specified in the framework. The soundness question underneath is mission breadth. Evaluation, standards-setting, assessor certification, interagency and international coordination, and original RSI measurement are not a large number of functions in budget terms, but each is scientifically immature and competes for the same scarce expertise, and a new institution would be building all of them at once. The risk is less the size of the mandate than the simultaneity of standing up several first-of-their-kind capabilities, which is a common way for new agencies to underdeliver.

*At the same time, policymakers should be realistic about the challenges of building a new institution. Their immediate priority should be developing CAISI's technical expertise, operational capacity, and institutional credibility. New responsibilities should be introduced incrementally and paired with the personnel, infrastructure, funding, and authorities necessary to execute them successfully. The effectiveness of any framework will ultimately depend on how it is designed, resourced, and executed in practice, and we expect those details to be refined through continued engagement among policymakers, industry, and technical experts. The recommendations that follow suggest guiding principles for institutional design rather than prescribing every implementation detail and propose a phased approach to building an institution capable of supporting policymakers while also maintaining the speed, rigor, and technical sophistication expected by the private sector.*

- *Build CAISI’s foundation. Before CAISI can take on significant new responsibilities, policymakers must ensure that it has the resources, authorities, and institutional support necessary to succeed. Policymakers should:*
- *Authorize CAISI and appropriate funding. Establish CAISI as a permanent institution with clear statutory authorities and sufficient funding to conduct frontier model evaluations, develop safety standards, certify third-party assessors, and coordinate with national security and scientific agencies, as well as with international partners.*

**Analysis**

The draft Obernolte-Trahan House bill implies funding at roughly $100M per year. The binding constraint at that level is less raw compute, since running evaluations is inference rather than training and is comparatively cheap, and more the cost of scarce expert personnel competing with industry

compensation, in addition to the cost of standing up classified computing environments. Limited funding on those fronts would make the deadline rule in item 3 harder to meet.

- *Elevate CAISI's authority and coordinate government support. The CAISI Director should report directly to the US Secretary of Commerce or another senior Cabinet-level official, and the White House should coordinate staffing, resources, expertise, and operational support from departments and agencies across the federal government.*
- *Provide flexible hiring authorities and public-service pathways. Adopt hiring authorities similar to those used by CHIPS for America, allowing CAISI to recruit technical talent quickly and offer competitive compensation, while creating pathways for experienced AI researchers, engineers, and safety experts to serve temporary tours of duty in government.*

**Analysis**

Building a credible frontier evaluator requires hiring people who have worked at frontier labs, because they have expertise that barely exists anywhere else. That makes a revolving door close to unavoidable, which is the very reason recusal rules, conflict-of-interest firewalls, and cooling-off periods matter. The framework proposes this hiring model, but does not mention specific guardrails. CHIPS-style authority also solves only the speed of hiring, not the pay gap with industry.

- *Mobilize national security expertise and data. Direct national security and scientific departments and agencies to support evaluations in these domains in coordination with CAISI and immediately make available personnel and data related to cyber, CBRN, and other national security domains to bolster CAISI's ability to assess and mitigate risks.*

**Analysis**

This section is where the "unique government capabilities" from the preamble would have to become operational. The blueprint frames it as internal to government: other national-security and scientific agencies detailing personnel and pushing classified cyber and CBRN data into CAISI's evaluations. The arrangements that do not yet exist are therefore the interagency ones: the clearances for CAISI staff to receive the material and the channels for agencies outside Commerce to feed a Commerce-housed evaluation, which is the same command-relationship gap flagged in item (ii).

- *Secure access to classified compute. CAISI should have access to classified computing environments capable of conducting frontier model evaluations, whether through dedicated infrastructure, interagency partnerships, or formal agreements with agencies or commercial providers. Create a mandatory evaluation process. Once CAISI has sufficient technical expertise and operational capacity, policymakers should require the most capable frontier models to undergo a CAISI evaluation before public release. These evaluations should assess frontier capabilities, the effectiveness of associated safeguards and mitigations, and the resulting risk profile of the deployed system. The requirement should be narrowly targeted at the most capable systems and should allow for iterative deployment by establishing clear thresholds for when subsequent model versions require reevaluation. CAISI's role should be to conduct evaluations and recommend mitigations—not to approve or block deployments. Developers should remain responsible for deployment decisions, publicly disclose evaluation findings and how they*

*responded, and remain accountable through transparency and reporting requirements. Policymakers should also ensure that evaluations are completed within a defined statutory timeline and that the evaluation process does not disincentivize the beneficial process of iterative deployment. If CAISI fails to complete an evaluation within the defined time period due to bandwidth, hardware, personnel, or other constraints, developers should be permitted to deploy without penalty. Companies should also remain free to share models with trusted third-party evaluators, independent researchers, red-teamers, and testing partners before or alongside CAISI review. A strong evaluation ecosystem requires multiple sources of expertise rather than a single institutional gatekeeper.*

**Analysis**

A real prerequisite for frontier evaluation, and a non-trivial procurement and security undertaking in its own right.

**Analysis**

This is the operational heart of the framework, and the place where "mandatory" quietly becomes closer to "advisory."

If an evaluator cannot withhold permission, this effectively is not really a gate, and in practice it works as a well-funded government red team wired to a disclosure requirement.

The phrase "remain accountable through transparency and reporting requirements" is carrying significant weight. A developer that disagrees with a CAISI finding may still deploy, and the accountability that remains is of two kinds, neither of which is a stop. Ex ante, the developer must disclose CAISI's findings and how it responded, so the discipline is reputational and informational. Ex post, the separate accountability provision preserves liability for severe harms. For most of the framework, that combination is workable. For the catastrophic tail, it is the weakest point in the design, because liability assessed after a loss-of-control or CBRN event does not undo the event, and disclosure does not prevent the deployment the evaluator warned against.

The deadline provision then removes even that friction at the exact moment CAISI is overloaded, so the regime binds least when capacity is thinnest. Read against the $100M resourcing above, this is the soundness crux of the document: a mandate whose failure mode is automatic permission inverts the usual precautionary logic, since the consequence of the regulator being too busy is that the model ships. To make the evaluation meaningful, funding should either be matched to evaluation throughout or matched to a different default on a missed deadline, such as a time-limited extension rather than an automatic green light. The design avoids putting a single government censor between a company and release, which is a legitimate liberal-institutional value. It also takes the only hard stop out of the system, which is a real safety cost.

*Support independent technical assessments. Deployment-triggered evaluations alone may not provide governments with sufficient visibility into how frontier capabilities evolve over time. Policymakers should therefore direct CAISI to develop standards for independent technical assessments, establish a certification process for qualified third-party assessors, and help build a broader ecosystem capable of conducting these reviews, including by using AI itself. Policymakers should also require frontier developers above specified capability thresholds to undergo periodic independent technical assessments conducted by CAISI-certified organizations. An initial priority for these assessments should be providing policymakers with ongoing visibility into progress toward RSI, highly capable internal deployments, frontier model*

*security, internal monitoring practices, and the effectiveness of associated safeguards. RSI may become the defining frontier safety and governance challenge of the coming decade, yet policymakers currently lack reliable ways to measure progress toward it or to understand its implications. CAISI should therefore work with frontier developers, academic researchers, national security agencies, and international partners to rapidly develop methodologies, benchmarks, and indicators for measuring RSI and assessing what governance works. Given the pace of AI development, we encourage CAISI to treat RSI as an urgent priority. Building on this work, CAISI should develop common standards for independent technical assessments and support the creation of a broader ecosystem capable of monitoring RSI progress and safeguards over time. Frontier developers should share appropriate RSI-related measurements with CAISI, and qualified third-party assessors should periodically evaluate those measurements using CAISI-developed methodologies. Technical assessment findings should be provided to CAISI, which should use this information to help policymakers understand progress, assess whether mitigations are keeping pace, and coordinate with national security agencies and international partners on the implications of AI progress and whether additional safeguards or policy responses may be warranted. Policymakers should also assign CAISI the resources and authorities necessary to help build the assessment ecosystem itself. CAISI should be authorized to provide grants, cooperative agreements, and other forms of support to emerging assessment organizations, academic centers, and technical institutions developing evaluation, technical assessment, and auditing capabilities.*

**Analysis**

Standing up a network of certified assessors reduces reliance on any single evaluator, but it inherits the volume question from the auditing provision: with only a handful of frontier labs, government-conducted, second-party assessment may be the more feasible path than building a certified private ecosystem from scratch.

The decision hiding inside this is who actually shapes the standards. Whoever defines the RSI benchmark defines what counts as dangerous. The document assigns formal authorship to CAISI, which is directed to produce “CAISI-developed methodologies,” but it also directs CAISI to develop them in collaboration with frontier developers and to rely on measurements the developers themselves supply, so industry shapes the substance of the benchmark even where it does not hold the pen.

This is the most durable point of leverage in the whole framework, more durable than the statute, because the measurement layer underneath a neutral law is what gives the law its real content. A neutral statute sitting on industry-authored benchmarks is, in effect, an industry regime. It is also the keystone dependency: the evaluations, the transparency reports, the thresholds, and the audits all presuppose this measurement, which is the least mature and most contested deliverable in the set, so if this piece is slow or contested, nothing upstream binds.

**Analysis**

This provision is probably necessary, since independent assessment capacity is thin today and will not materialize on its own, with the standard caveat that a grant-making body shapes which methods and assessors grow, so the criteria are worth watching for a quiet preference toward industry-friendly approaches.

Read in sequence, this is really an admission about build order. The audit requirement and the third-party assessment regime both depend on an ecosystem this item exists to create, and that creation takes years, so although it is presented as a supporting detail, it belongs early in any realistic timeline. It also reinforces the prior point: if the ecosystem has to be built and funded from zero, the

second-party government assessment option deserves a serious look as the faster route at current scale.

## 3. Mobilizing a whole-of-government resilience strategy

*Strong institutions are necessary, but they are not sufficient. No evaluation process, assessment regime, or single organization can eliminate every risk. As AI becomes increasingly capable, democratic societies will need to make sure that defensive capabilities, public institutions, and societal resilience improve alongside them. Frontier AI should therefore be treated as a national priority requiring coordination across national security, public health, cybersecurity, scientific, diplomatic, and economic agencies, as well as with international partners. Policymakers should pursue a resilience strategy that not only reduces risk, but also strengthens society's capacity to respond to emerging challenges, adapt to changing conditions, and continue benefiting from AI.*

*Facilitate collaboration and international coordination on frontier AI safety. Frontier risks will not be addressed by any one organization or country acting alone. Policymakers should provide legal certainty that allows frontier developers to collaborate on safety-related issues, including sharing threat intelligence, evaluation methodologies, incident learnings, and best practices. Democratic nations should also work together to develop compatible safety frameworks, trusted channels for information sharing, and coordinated responses to serious incidents. Particular priority should be given to developing shared approaches for evaluating and responsibly communicating progress toward RSI, where a lack of shared measurements and transparency could intensify competitive pressures among developers and make it more difficult to determine when additional safeguards are warranted. The emerging network of AI safety institutes could help build this shared technical understanding and provide a foundation for coordinated action if additional safeguards become necessary. Over time, the network could help build the shared methods, technical expertise, and confidence-building measures needed for governments to assess whether agreed safeguards are being implemented effectively. Protect America's compute advantage. Advanced AI capabilities depend on access to leading-edge semiconductors and large-scale computing infrastructure. Policymakers should strengthen export controls, close known loopholes, and invest in the compute, energy, and infrastructure needed to maintain US leadership. Strategic compute capacity would give the US the ability to evaluate, govern, and deploy frontier AI systems when national security demands it. Maintaining leadership in advanced compute is not only an economic and national security priority—it's also a frontier safety strategy.*

**Analysis**

In plain terms this is a request for an antitrust safe harbor. When direct competitors share information, the blueprint is requesting assurance that safety coordination will not be treated as collusion.

There is precedent for this, but that precedent is narrower than it may look at first glance. Cybersecurity threat-sharing safe harbors are well established, but what they typically cover are indicators of compromise: narrow, factual, time-sensitive threat data. Sharing evaluation methodologies, best practices, and release-relevant learnings is a different and more competitively sensitive category, closer to coordinating how the industry operates than to swapping threat signatures. In other words, the precedent supports the threat intelligence piece and does much less work for the broader categories the document bundles in, which is exactly where the antitrust risk lives.

The hazard that this presents is that the same legal channel can be used to coordinate on release norms and capability thresholds, which softens competition and lifts barriers to entry. The competitive side effect is something that agencies tasked with reviewing antitrust should weigh directly, with the

permitted scope of sharing drawn narrowly, rather than something a safety statute grants by implication.

**Analysis**

This is the most independently operable item in the section, because the machinery already exists in Bureau of Industry and Security (BIS) export controls and existing compute and energy programs and it depends on no new standard. The flip side is that it endorses a track already in motion more than it proposes a buildable new mechanism, so its marginal feasibility cost is low, but so is its marginal content.

*Restrict the adoption of unevaluated frontier AI systems. Federal agencies should prohibit the use of frontier AI systems that have not undergone a recognized safety evaluation on government-owned systems and devices. These evaluations should assess systems as deployed, including associated safeguards and controls, rather than the underlying model in isolation. Agencies should also prohibit procurement of products and services that rely on unevaluated frontier models in sensitive government contexts. Trusted evaluation processes are essential to ensuring that frontier systems deployed within government meet appropriate security and safety standards. Ensure defensive capabilities scale faster than offensive capabilities. Frontier AI will strengthen both defenders and attackers. Policymakers should invest in AI-enabled biodefense, cybersecurity, critical infrastructure protection, and rapid response systems that reduce the consequences of misuse regardless of where threats originate. OpenAI's Cyber Action Plan provides examples of how advanced AI can strengthen resilience by helping trusted defenders identify threats earlier, respond faster, and protect critical systems more effectively. AI should strengthen the institutions and systems that protect society, ensuring that resilience grows alongside capability.*

**Analysis**

This provision is very reasonable at face-value – government should not run untested frontier models on its own systems.

This is one of the more enforceable levers in the document, because procurement is a tool the government fully controls and needs no new authority to use. That cuts both ways: if a recognized evaluation existed and were accessible, this could move quickly, but as written, it inherits the missing-standard problem and the access problem as described above, and it could imply instructing the government to buy from a short list of incumbents.

**Analysis**

I have two caveats on the merits of this provision. The idea that defense naturally outpaces offense is an assumption rather than a rule, and in some areas, including parts of biological risk, offense may move faster, so it should not be treated as a settled assumption. The operable part of this provision is specific program investments, which existing agencies and appropriations can deliver, whereas "ensure defense scales faster than offense" is a goal with no metric, which means no one can manage it or verify it.

*Prepare for future resilience challenges. Frontier AI governance will continue evolving as capabilities advance. In Industrial Policy for the Intelligence Age, we outlined several potential approaches to strengthening societal resilience, including AI trust infrastructure, model-containment playbooks, safety systems for emerging cyber and biological risks, and new approaches to international coordination. Policymakers should direct relevant agencies to evaluate the feasibility, costs, benefits, and implementation challenges associated with these and other proposals, and identify where additional authorities or resources may be required. Building resilient institutions today will help democratic societies adapt as future governance challenges emerge.*

## Building the institutions for democratic governance

*Frontier AI will help shape the balance of economic, scientific, and geopolitical power in the 21st century. Democratic societies have a narrowing window to build the institutions needed to govern increasingly capable AI systems before frontier capabilities outpace existing frameworks. States have helped establish an emerging frontier safety baseline. Policymakers should now build on that foundation by creating a national framework, strengthening CAISI as the US federal government's primary frontier AI institution, and mobilizing a broader resilience strategy. Because frontier AI is a global technology, democratic nations must also work together to strengthen technical institutions, develop compatible governance approaches, and coordinate responses to emerging risks. The framework outlined here is not intended to be the final word on frontier AI governance. AI is advancing rapidly, and many important questions remain unresolved. The goal is to build the institutions, standards, and resilience needed for democratic societies to understand, adapt to, and govern increasingly capable AI systems.*

**Analysis**

The urgency is a real concern, and it also creates a challenge for Congress. A fully drafted architecture arriving on an urgency timeline invites adoption of the ready-made version rather than the slower legislative work of weighing stakeholders and calibrating the specifics. The structural pieces, preemption and institutional design above all, are the hardest to revisit once they are law, so the cost of moving quickly falls heaviest exactly where reversibility is lowest.

The structural pieces, especially preemption and institutional design, are the hardest things to revisit once they are law.

**Analysis**

There is also a tension between the urgency and the document’s own architecture. The framework cannot in fact be stood up quickly, because its core deliverables—the RSI and loss-of-control measurements, the evaluation methodologies, and the assessor ecosystem—take years to produce. So even if Congress acted tomorrow, the honest timeline to a functioning regime is long, and the genuine risk is a sequencing mismatch: enacting the fast, structural pieces (preemption above all) on the urgency timeline while the slow, substantive pieces lag, leaving a period in which state law is switched off and the federal framework cannot yet bind.

# Appendix 1

**Frontier safety frameworks (the major labs)**

1. OpenAI, *Preparedness Framework (Version 2)*, April 15, 2025, https://openai.com/index/updating-our-preparedness-framework/.
2. Anthropic, *Responsible Scaling Policy* (Version 3.1, effective April 2, 2026), https://www.anthropic.com/responsible-scaling-policy.
3. Google DeepMind, *Frontier Safety Framework (Version 3.0)*, September 22, 2025, https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf.

**Transparency-report instruments (the per-model disclosures these frameworks produce)**

1. Anthropic, *Responsible Scaling Policy* (Version 3.1), sections on Frontier Safety Roadmap and Risk Reports, https://www.anthropic.com/responsible-scaling-policy.
2. Frontier Model Forum, *Issue Brief: Components of Frontier AI Safety Frameworks*, November 8, 2024, https://www.frontiermodelforum.org/updates/issue-brief-components-of-frontier-ai-safety-frameworks/.

**Statutory and international disclosure schemas**

1. California Senate Bill 53, *Transparency in Frontier Artificial Intelligence Act* (TFAIA), signed September 29, 2025, key provisions effective January 1, 2026, official text at https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53
2. European Commission, *General-Purpose AI Code of Practice*, final version published July 10, 2025, https://digital-strategy.ec.europa.eu/en/news/general-purpose-ai-code-practice-now-available; full text at https://code-of-practice.ai/

**Other federal frameworks by major AI Labs:**

1. June 2026: Anthropic, *Advanced AI Framework*, https://www-cdn.anthropic.com/files/4zrzovbb/website/0a58d567024a8b448ff15158ebc3625328dfcc1f.pdf